Low Resource Offline Multimodal Chatbot for ScienceQA

David Liu, Jazmin Guo, Agrima Bansal, Shihao Xiong, Vansh Pruthi

***Abstract*—Large language models (LLM's) are increasingly being adopted to assist students with academic tasks; however, exposing children to continuous internet access raises significant concerns regarding accessibility, privacy, and safety. To address these issues, this project introduces an efficient science chatbot tailored for school students, designed to operate directly on local devices without requiring internet connectivity or substantial computational resources. Our system integrates a compact language model (Gemma-3-4b) with a retrieval-augmented generation (RAG) framework that selectively activates retrieval based on the model's confidence, utilizing ChromaDB and MiniLM-L12-v2 to enhance answer quality. To improve accuracy while maintaining low resource usage, we employ a combination of fine-tuning, RAG, and model quantization. To further improve efficiency, a confidence-based mechanism determines when to retrieve supplementary information to support the model's responses. Experiments demonstrate that using a mixed methods approach, our fine-tuned and quantized model achieves 73.52% accuracy on the ScienceQA dataset, presenting a viable solution for educational settings with limited resources.**

Keywords— Offline Chatbot, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Confidence, Log Probability, Vector Database, ScienceQA

I. INTRODUCTION

This project aims to design and implement a local, low resource chatbot powered by a large language model (LLM) to accurately answer primary and high school science questions. While popular state of the art LLMs such as ChatGPT and Gemini demonstrate high performance, they rely on continuous internet and expensive subscription fees, making them impractical for many students. We want to improve upon existing proprietary LLMs by designing a local LLM with adaptive processing based on LLM token confidence.

We propose a more accessible alternative: an offline chatbot that operates on personal devices using the Gemma-3-4b-it open-source model [4]. Our approach uses the LLM's token-level confidence to determine when retrieval is necessary. High-confidence predictions are returned directly, while low-confidence ones trigger a retrieval-augmented generation (RAG) step using ChromaDB and MiniLM-L12-v2 to fetch relevant science content. To further optimize performance, the instruction tuned Gemma-3-4b model was further fine-tuned using LoRA, significantly reducing the number of trainable parameters and improving model efficiency. Additionally, we applied quantization techniques to minimize computational load and memory usage, enabling the system to run smoothly on devices with limited resources such as laptops and smartphones.

These methodologies combined balances efficiency and accuracy, reducing computational costs without sacrificing performance

II. RELATED WORK

The ScienceQA dataset introduced by Lu et al. (2022) serves as a major benchmark for evaluating science question-answering systems. It contains over 21,000 multimodal multiple-choice questions that cover science topics in various fields and are suitable for school-level education. They also introduced a key feature, Chain-of-Thought (CoT) reasoning, a prompting technique that allows models to generate answers through multiple reasoning steps. CoT was shown to improve accuracy on complex questions, particularly those requiring multi-hop reasoning [6]. However, their research mainly focused on large-scale and cloud-based models like GPT-3, it needs to be refined for offline and low-resource environments.

Lewis et al. (2020) proposed Retrieval-Augmented Generation (RAG), a framework that combines a pretrained language model with an external retriever. This framework is aimed to address factual correctness in models. RAG can retrieve relevant documents from external resources (like Wikipedia) in the model's generation. This architecture greatly enhances the factual accuracy of language models on tasks that contain intensive knowledge [2,7]. However, RAG implementations require large retrievers and models, which limits their usability on smaller devices and devices do not support receiving external documents [2].

This research work shows the technologies that serve as the foundation of our project. We acknowledge the importance of reasoning, retrieval, and confidence calibration in improving QA systems [8, 9]. However, we found these articles were assuming powerful computation devices and well-connected internet. To address the gap, we consider balancing performance and efficiency by applying these technologies within a small, offline model (Gemma 3-4B), making educational AI tools more accessible.

III. METHODS

This project employs a combination of methods to balance accuracy and resource efficiency for an offline science question-answering chatbot. We selected the Gemma 3-4B instruction-tuned model as the core language model due to its relatively small size, allowing it to run on local devices while providing considerable and reasonable performance [4]. This model is pre-trained and further optimized through fine-tuning on the ScienceQA dataset, which contains over 21,000 science questions across various topics and difficulty levels.

To improve factual accuracy and handle complex questions, we integrate Retrieval-Augmented Generation (RAG). This method retrieves relevant external documents from a ChromaDB vector database, a lightweight vector database designed for embedding-based search. For this, we use MiniLM-L12-v2 embeddings due to their balance between speed and semantic accuracy. The retriever identifies top-k relevant documents based on cosine similarity, which are then included in the model's prompt when generating answers. However, to reduce computational overhead, RAG is only triggered when the model's confidence score is low, following a confidence-based adaptive computation strategy. This allows the chatbot to respond directly when confident and use additional context when uncertain.

Fine-tuning is applied using LoRA (Low-Rank Adaptation), a method that inserts trainable low-rank matrices into transformer layers, greatly reducing computational overhead. LoRa is applied within the PEFT (Parameter-Efficient Fine-Tuning) framework, the framework enables selective tuning of parts of large models. This approach adjusts only specific parts of the model (rank matrices), significantly reducing the training cost compared to full fine-tuning. Fine-tuning the model on the ScienceQA training set enables the chatbot to adapt more closely to the domain-specific terminology and reasoning patterns.

Additionally, since most students' device is lower than 8 GB, quantization is adopted to compress the model to 8-bit integer precision, reducing memory usage and increasing inference. This ensures the model can run on laptops or smartphones.

By applying fine-tuning, RAG, confidence-based retrieval, and quantization together, our system optimizes the trade-off between accuracy and resource constraints, making it practical for offline educational environments. This strategy enhances both the efficiency and effectiveness of science question answering.

IV. EXPERIMENTAL SETUP

A. DATASET SPLIT

The ScienceQA dataset consists of 21,208 multimodal multiple-choice questions that are randomly split into training, validation, and test splits with a ratio of 60:20:20. There are approximately 12,725 questions allocated for training, 4,242 for validation, and 4,241 for testing.

The dataset spans 26 subject areas such as physics, chemistry, history, grammar and vocabulary that span from grade 1 to grade 12 difficulty levels.

B. EVALUATION STRATEGY

We evaluate the model's performance by examining how different combinations of methods—such as RAG, mixed methods, fine-tuning, and quantization—affect both accuracy and efficiency. Specifically, we assess how these methods contribute to improvements in the correctness of the selected options and overall model performance.

C. MODEL HYPERPARAMETERS

The entire setup is run on google colab with the following base configurations:

1. GPU: A100 for fine tuning and L4 for tests
2. Base Model: We make use of Gemma3-4B-IT which is optimized for multimodal activities and facilitates effective question-answering in a variety of topics.
3. Embedding Model: The Sentence-Transformers all-MiniLM-L12-v2 model is used for generating dense vector embeddings of text.
4. Vector Database: ChromaDB to store and retrieve embeddings. It utilizes cosine similarity for top-k=1 configuration to retrieve the most relevant context for each query.
5. Fine-tuning: The model is fine-tuned using Low-Rank Adaptation (LoRA) to reduce the computational cost of fine-tuning the full 4B parameters. Initially, Base64 encoding is used for image prompts, which is replaced by URL-based image prompts to retain more detailed information.
6. Quantization: The model is quantized to 8-bit precision, significantly reducing memory usage.

V. RESULTS

Our experiments evaluated various methods such as zero-shot prompting, Retrieval Augmented Generation (RAG), fine-tuning, quantization and mixed methods (RAG based on confidence) on the ScienceQA dataset.

A. FINE-TUNING RESULTS

base64 image prompting	
Base64 string Pretrained (quantized)	59.5
Base64 string Fine-tuned (quantized)*	65.8

Table 1 Fine-tuning results with base64 Prompting (1000 test samples)

In this setting, the results were obtained by passing the images in the questions as base64 strings in the prompt. This method of handling multimodal inputs significantly lowers accuracy, since most of the input was occupied by the image string, the weight of the literal question is low. This Evaluation was conducted on a randomly selected 1000 samples from the test set instead of the full ~4k samples.

Fine-tuning the dataset resulted in a 6.3% increase in accuracy, demonstrating effective model optimization. Further improvement was achieved by using URL image prompts instead of base64, resulting in a 5.72% accuracy gain (as shown in Table 2, parameter with *). While the improvement over the zero-shot baseline was not substantial, these results indicate that even with the image handling challenges, the model's performance was enhanced through the fine-tuning process.

B. METHOD ACCURACY COMPARISON

Model Setup	Accuracy (%)
URL Fine-tuned (quantized)*	71.52
Zero-shot (quantized)	73.21
RAG (quantized)	65.64
RAG (URL fine-tuned + quantized)	63.36
Mixed method ~ Full test set (URL fine-tuned + quantized)	**73.52**

Table 2 results with full test set and URL case prompting for finetuning (when mentioned)

The zero-shot performance of the Gemma 3-4B model reached 73.21% accuracy, which is quite strong considering the model's small size and quantized format. This shows that even without fine-tuning, the model can handle basic science questions effectively.

RAG alone lowered accuracy to 65.64%, indicating that external context doesn't always improve answers for smaller models, possibly due to difficulties in processing long documents. Fine-tuning and quantization slightly reduced accuracy to 63.36%, but significantly improved efficiency and processing time.

The mixed method of fine-tuning with quantization achieved 73.52% accuracy on full test set. A 0.31% improvement over zero-shot. Though the improvement is small, the mixed method shows potential ability for adaptive computation, meaning the system uses retrieval only when needed for more efficient and reliable offline performance. This result took 25 minutes and 41 seconds to run the 4241 questions, meaning the model took about 0.36 seconds to answer each question. This means our model is extremely computationally efficient.

Overall, these findings indicate that Gemma 3-4B, even in its quantized form, performs well as expected for offline science question answering. The fine-tuning and RAG offer slight improvements, the zero-shot model's strong baseline suggests that small models can perform well without extensive adaptation, making them suitable for low-resource, real-world applications. The mixed method enhances system stability by applying retrieval selectively, balanced performance and efficiency.

C. MODEL ACCURACY COMPARISON

We extracted a difficult subset of the questions that GPT 4o-mini got wrong. Then we used this subset to evaluate our strategies.

Model	Accuracy (%)
GPT-4o Zero-shot	49.28
Gemini-2.0-Flash Zero-shot	55.21
Gemini-2.0-Flash with RAG (#1)	60.82
Gemini-2.0-Flash with RAG (#2)	57.96

Table 3 results from the “difficult subset” because it was expensive to run on full test set.

Model	Accuracy (%)
Gemma3-4B Zero-Shot Quantized	49.28
Gemma3-4B Zero-Shot (Fine-tuned + Quantized)	55.21

Table 4 results from full test set

The Gemma 3-4B model, although size and parameters are much smaller than models like GPT-4o or Gemini-Flash, achieved 73.31% accuracy on the relatively difficult questions sets when fine-tuned and quantized. This performance showing our approach is powerful to specified domain knowledge.

We concede that popular models such as GPT-4o and Gemini-Flash typically perform better on general and broader datasets, however, they rely on online access and costly subscription. Even in some cases, they require latest system software support, such as only support Android or iOS version in recent years. This make those devices is not purchased in recent or unable to afford new devices, inaccessible to the two models. In contrast, our model offers better performance while running entirely offline and have more adoptability to hard question reasoning and answering, making it suitable for education -level use.

VI. CONCLUSION

In this project, we developed an offline, low resource chatbot that answers science questions from primary to high school levels using the ScienceQA dataset. Our main contributions include using the Gemma3-4b model with a combination of techniques like fine tuning, quantization and confidence-based RAG. These approaches enable the chatbot to run efficiently while keeping the computational cost low so that they can be run on local devices like laptops and mobile phones.

The key advantage of this project was that the chatbot runs efficiently on educational questions and doesn't need internet access and is completely free addressing both privacy concerns, hardware concerns and educational equity.

The key advantage of our solution is its ability to achieve competitive accuracy (73.52%) in offline environments, balancing performance and resource requirement.

However, the current study has some limitations. The Gemma 3-4B model, being relatively small, is hard to work with long retrieved documents, which limits the effectiveness of RAG. Additionally, the fine-tuning improvements were also limited, likely due to the model's capacity constraints.

Given more time and resources, several improvements could be made:
1. Integrating context summarization to help the model process retrieve documents more effectively.
2. Exploring knowledge distillation from larger models to improve the small model's performance without increasing its size.
3. Explore the impacts on performance from quantization and gguf conversion when moving to mobile devices.

These enhancements would further strengthen the chatbot's performance and expand its applicability, making it an even more valuable tool for offline educational environments.

References

[1] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., & Kalyan, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. arXiv. https://doi.org/10.48550/arXiv.2209.09513

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. https://doi.org/10.48550/arXiv.2005.11401

[3] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2106.09685

[4] Gemma Team. (2025). Gemma 3 Technical Report. arXiv. https://doi.org/10.48550/arXiv.2503.19786

[5] Github.io. (2022). ScienceQA: Science Question Answering. [online] Available at: https://scienceqa.github.io/#dataset

[6] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. and Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs]. [online] Available at: https://arxiv.org/abs/2201.11903

[7] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J. and Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. [online] arXiv.org. Doi https://doi.org/10.48550/arXiv.2312.10997

[8] Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P. and Gurevych, I. (2023). A Survey of Confidence Estimation and Calibration in Large Language Models. [online] arXiv.org. Available at: https://arxiv.org/abs/2311.08298

[9] LeCoz, A., Herbin, S. and Adjed, F. (2024). Confidence Calibration of Classifiers with Many Classes. [online] arXiv.org. Available at: https://arxiv.org/abs/2411.02988